Nowcasting and Forecasting COVID-19 Cases and Deaths Using Twitter Sentiment

David Askay
California Polytechnic State University
daskay@calpoly.edu

Declan Molony
California Polytechnic State University
dmolony@calpoly.edu

Hunter Glanz
California Polytechnic State University
hglanz@calpoly.edu

Julia Alber
California Polytechnic State University
jmalber@calpoly.edu

Abstract

Real-time access to information during a pandemic is crucial for mobilizing a response. A sentiment analysis of Twitter posts from the first 90 days of the COVID-19 pandemic was conducted. In particular, 2 million English tweets were collected from users in the United States that contained the word 'covid' between January 1, 2020 and March 31, 2020. Sentiments were used to model the new case and death counts using data from this time. The results of linear regression and k-nearest neighbors indicate that sentiments expressed on social media accurately predict both same-day and near future counts of both COVID-19 cases and deaths. Public health officials can use this knowledge to assist in responding to adverse public health events. Additionally, implications for future research and theorizing of social media's impact on health behaviors are discussed.

1. Introduction

The Coronavirus Disease 2019 (COVID-19) has contributed to nearly 661,000 deaths in the United States (US) since February 2020. It is essential to have access to timely and accurate information for coordinating resources and making policy decisions during public health events, such as the COVID-19 pandemic. In the US, the National Notifiable Diseases Surveillance System monitors infectious diseases through integrating data from a network of health departments, laboratories, and hospitals across local, tribal, territorial, state, and federal agencies. While these reports are of high quality, the time to collect and verify this data can slow down the publication and use of these critical reports. Faced with the challenges of reporting COVID-19 related data, the CDC cites an urgent need to modernize its disease surveillance infrastructure to create access to real-time health data—"data that moves faster than disease" [1].

Coinciding with this need for real-time access to data are scholarly efforts to address this need through data on the internet. Analysis of social media content and user behavior has successfully been used during prior outbreaks to monitor both public sentiments and spread of diseases [2]. Indeed, Gunther Eysenbach in 2006 used Google Search trends to predict influenza epidemics more accurately and timelier than relying on traditional methods of reporting [3]. He subsequently proposed the concept of *infodemiology*—the analysis of internet information and communication patterns during public health events to predict and inform public health professionals and policymakers [4, 5]. From analyzing trends in internet searches to the volume and sentiment of social media posts, scholars are actively developing procedures for generating real-time information on public health events that can be used for early detection, predicting spread, and informing responses [6].

Unsurprisingly, scholarly attention on infodemiology has flourished throughout the COVID-19 pandemic, with researchers investigating various aspects of social media to predict cases and track sentiments associated with the disease. For example, Google searches for coronavirus correlated with case and death rates around the world [7]. An analysis of the volume of COVID-related tweets generated each day in Italy predicted the total number of deaths a month later [8]. Similarly, an analysis of the Sina Weibo platform, often seen as China's equivalent to Twitter, found that the volume of COVID-19 related for posts positively correlated to the number of reported case rates in Wuhan [9].

URI: https://hdl.handle.net/10125/79850
978-0-9981331-5-7

The present study contributes to this body of research by investigating the degree to which Twitter posts can be used to provide same-day (nowcasting) and predictive (forecasting) information on COVID-19 case rates and death rates in the US. The unsupervised algorithm *vaderSentiment* [10] was used to determine the sentiment of 2 million tweets containing the word 'covid' from between January and March 2020. Additionally, each post's meta-data, consisting of the number of likes, number of followers, whether the tweet sender was verified, and number of retweets, was also included in the analysis. A prediction model using tweet sentiment was explored to estimate COVID-19 case and death counts. The results of a regression analysis indicate that average tweet sentiment at the state level may be predictive of COVID-19 cases and deaths one and two days into the future.

2. Literature Review

Social media provides people a means to express and react to real-time opinions, thoughts, emotions, attitudes, and behaviors. Additionally, posts offer valuable meta-data that provide information like geographical location of the poster, reach of the poster (e.g., number of followers), credibility (e.g., verified user accounts), public response (e.g., number of likes and reposts). This allows researchers to localize and aggregate data to understand people's reactions, awareness, outlooks, and beliefs concerning events. To public health professionals, sentiments expressed on social media provide valuable insights for understanding and responding to emergent public health needs [11]. For example, studies have found significant relationships between the volume of disease-relevant posts and case counts for influenza [12], the zika outbreak [13], and the avian flu [14].

Used by millions of citizens, politicians, and government institutions, Twitter is among the most used social media platforms used for sharing and acquiring information about global adverse events. This has been particularly true for the current COVID-19 pandemic [15]. Since Twitter data is available in real-time, it has the potential to overcome delays in reporting and offer immediate insight to public health officials and policymakers. As a result, scholars have turned to Twitter to investigate topics such as monitoring public perceptions and misunderstandings [16], early detection of outbreaks [17], and the spread of misinformation [18].

Sentiment analysis is a common approach to analyzing Twitter posts. This approach applies natural language processing to classify text along a spectrum as having a negative, neutral, or positive connotation. While a post expressing enthusiasm for receiving a vaccine would likely have a positive connotation, a post expressing concern for a family member diagnosed with COVID-19 would fall more on the negative spectrum. At the aggregate level, sentiment analysis offers insight into Twitter user's real-time and dynamic emotional state. This makes sentiment analysis of tweets a promising technique for forecasting, having been used to predict election outcomes [19], stock market movements [20], and physical activity levels [21]. Moreover, the accuracy of models predicting cases of the 2009 H1N1 outbreak was improved by including the sentiment of social media posts [22].

Applying sentiment analysis to Twitter can provide valuable information for combating outbreaks and pandemics [23]. Several studies have already applied sentiment analysis to Twitter posts during the COVID-19 pandemic to compare sentiment across countries [24] and learn sentiment towards specific topics (e.g., wearing masks) [25]. While it has been proposed that sentiment analysis may be useful for predicting case and death counts for COVID-19 [11], little published research currently exists.

Filling this gap in the literature, this study draws from 90 days of US-based Twitter posts to investigate the potential of sentiment analysis of posts to predict case and death counts of COVID-19.

3. Research Design and Methodology

3.1. Data Collection and Preprocessing

Similar to previously used infodemiology methods [26], this study focuses on Twitter data and machine learning approaches to generate real-time and future insights into public health concerns. Twitter data was collected using a Python library called *snscrape*. This web scraper collects the tweet content and meta-data (e.g., date and time, user location, number of likes, number of followers, number of retweets, and whether the sender is verified) for publicly available Twitter posts. We collected over 10 million tweets from between 01-01-20 and 03-31-20 containing the word 'covid.' Of this collection of tweets about 2 million were from users in the United States and in English, which were included for subsequent analysis.

The Python library *nltk* was first used to remove stopwords from each Tweet. Stopwords are common words such as 'the' or 'and' that do not convey any particular connotation or emotion. These words were removed, along with any non-alpha characters, for the sentimental analysis. Since the prevalence of COVID-19 differed by location, Twitter posts were also geocoded by the state of the Twitter user account. This allows us to investigate the impact of localized Twitter posts on the case and death counts in each US state.

COVID-19 case and death counts per day were acquired from the open-source repository maintained by the New York Times on GitHub [27]. This repository compiled outbreak data on a daily level from state and local governments, and multiple United States health departments. Historical data for cases and death counts were collected for the time period under investigation. In this study, we are interested in exploring the relationship between sentiments and both same-day and future case and death counts.

3.2. Sentiment Analysis

The sentiment analysis of tweets was conducted using the Python library *vaderSentiment* [10]. Public health researchers have successfully used this algorithm to investigate COVID-related sentiments expressed on Twitter [28]. The Valence Aware Dictionary for sEntiment Reasoning (VADER) is a lexicon and rule-based model for general sentiment analysis of social media. VADER assigns each Tweet a score from -1 (negative connotation) to 1 (positive connotation) with 0 representing neutral. It works by assigning a sentiment value to text based on a lexicon of over 7,500 words, abbreviations, emoticons, and jargon (including slang like '*sux*' and exaggerations like '*uber*' and '*friggin*') while also considering sentence structure, negation, and hedging to gauge the intensity of the emotion (e.g., good vs. very good). Despite these documented strengths, VADER is also limited in its ability to completely account for instances of sarcasm or multi-polarity, as is the case of many sentiment analysis approaches.

The sentiment score, aggregated across users in different states, represents the emotional response to the pandemic of Twitter users in the US. Figures 1 and 2 show a word cloud for the top 100 frequently used words in positive (>.5) and negative (<-.5) tweets.

Figure 1. Word cloud of positive sentiment tweets

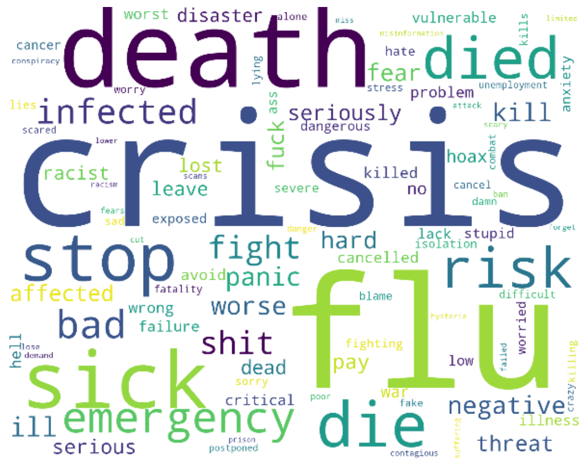

Figure 2. Word cloud of negative sentiment tweets

Finally, we are interested in exploring the relationship between sentiments of social media and both same-day and future case and death counts in the United States. The sentiment score allows for testing same-day relationships. To investigate if using current social media data can predict future cases and deaths, lagged sentiment variables were created that lagged the sentiment variable by one to five days. This allows us to explore the extent to which the average sentiment on one day can predict cases and death on the following five days. All the numeric variables (i.e., death count, case count, sentiment scores, lagged sentiment scores) were averaged after grouping by state and date to proxy a state's reaction. Our analysis considers the entirety of this data from all 50 states on each day in a single model.

3.3. Regression Analysis

A regression analysis was conducted to investigate the relationships between social media sentiments and COVID-19 cases and deaths. Additional variables were considered in this analysis, using metadata from each tweet. These variables include the number of likes, number of followers, whether the sender is verified, and retweet count. To avoid overfitting the model, variable selection was considered through forward and backward selection of the variables presented in Table 1.

Through comparing Bayesian information criterion (BIC) scores, forward and backward selection identified the most important variables to be *sentiment score*, *time*, and *state*. However, using the lagged data sets revealed the lagged versions of sentiment score were more important than the real-time sentiment score in predicting cases and deaths. This suggests that statements on Twitter might be predictive of future COVID-19 case and death rates. Based on the variable selection process, the subsequent regression analysis included the variables identified in Table 2. Of note, the sentiment lagged by one day was the most important for predicting future cases, while the sentiment lagged by two days was most important for predicting future deaths.

Table 1. Variables used for the selection procedure

Variable	**Description**
Sentiment Score	A score between -1 and 1
Lagged Versions of Sentiment Score	Five versions of sentiment score were created, each one lagged by an additional day
Number of Likes	How many likes the Tweet received
Number of Followers	How many followers the Tweet sender has
Verified	Whether the Tweet sender is verified on Twitter (i.e., blue check-mark)
Retweet Count	How many times a Tweet was retweeted
Time	The date expressed as a fractional year
State	Each observation is one of the fifty states

Two regression methods were used to predict case and death counts: linear regression and K-Nearest Neighbors (KNN). Using cross-validation, test error estimates were approximated and compared. Cross-validation is a method to evaluate the performance of models whereby the data is separated into two subsets: a training dataset and a test dataset. We used 5-fold cross-validation, meaning that the dataset was split up into 5 subsets and each one takes a turn being the test dataset, and then the test error estimates are averaged.

Several transformations were applied and considered with linear regression. Polynomial versions of the time variable were experimented with since COVID-19 grew exponentially. However, when polynomial versions of time were included in models, R-squared only increased at most by 1.5%, which we deemed not a sufficient enough increase to warrant its inclusion.

Table 2. Variables included in the regression analysis

Predicting Cases with same-day data set	Sentiment Score, Time, State
Predicting Cases with lagged data set	Lagged by one day Sentiment Score, Time, State
Predicting Deaths with same-day data set	Sentiment Score, Time, State
Predicting Deaths with lagged data set	Lagged by two days Sentiment Score, Time, State

A mixture of two regularization techniques, LASSO [29] and Ridge Regression [30], was used to see which variables were most important and enhance prediction accuracy. When these two regularization techniques were applied on the regular and lagged data sets, they also identified and selected the same variables that were found to be important in the previous forward and backward selection procedures. When predicting cases or deaths, the coefficient on Sentiment Score was negative. This means that, on average, positive tweets about COVID-19 (e.g. "*staying at home saves lives*"), were associated with lower numbers of cases and deaths. In contrast, negative tweets, on average, about COVID-19 (e.g., "*This whole covid scare is a stupid hoax and is destroying this country!*") were associated with higher numbers of cases and deaths.

With KNN, a grid of values was hypertuned to find the best number of neighbors on cross-validated data. A more flexible model such as KNN might be better for prediction because linear regression assumes a parametric form and that may be too strong of an assumption for this context.

4. Results

To predict cases and deaths as accurately as possible, multiple models were explored. These include linear regression, regularization of linear regression, and

KNN. Table 3 and Table 4 show the R-squared estimates of these models on the test data predicting cases and deaths using the same-day and lagged sentiment data, respectively.

Table 3. R-squared test estimates for same day cases and deaths

	Linear Regression	Regularization of Linear Regression	KNN
Same Day Cases	0.3312	0.3285	0.9830
Same Day Death	0.2853	0.2854	0.9722

Table 4. R-squared test estimates for lagged cases and deaths

	Linear Regression	Regularization of Linear Regression	KNN
Lagged Cases	0.3780	0.3794	0.9648
Lagged Death	0.3204	0.3260	0.9551

The lagged versions of linear regression outperformed the real-time versions when comparing R-squared. This suggests there may be a relationship between current social media sentiment and future cases and deaths. Overall, the best model was predicting same-day COVID-19 cases with KNN. With that model, 98.3% of the variation of cases can be explained by its relationship with the predictors. Still, the remaining models were able to predict same-day or lagged cases and death rates at 95% or higher.

Figure 3 shows a heatmap of the logs of the total deaths over the period of the study, and Figure 4 shows a heatmap of the logs of the total cases over the period of the study. Logs were used for these figures to offer better color separation between states. Figure 5 shows a heatmap of the average tweet sentiment over the study period.

The high predictive accuracy of KNN is attributed to its capacity to better capture the complexities of this data compared to traditional linear regression modeling. Linear regression was most likely hindered by the strict assumption of linearity between the response and explanatory variables. These results suggest that the KNN model generated on the training data would perform similarly on a subsequent novel test dataset, and indeed better than a traditional linear regression model.

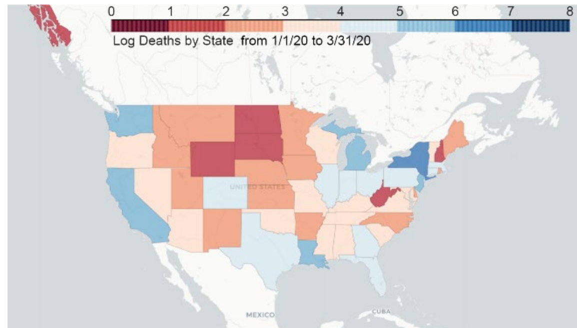

Figure 3. Heatmap of log of the total deaths by state from 1/1/20 to 3/31/20

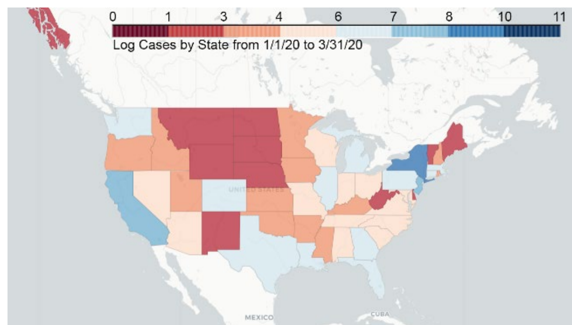

Figure 4. Heatmap of log of the total cases by state from 1/1/20 to 3/31/20

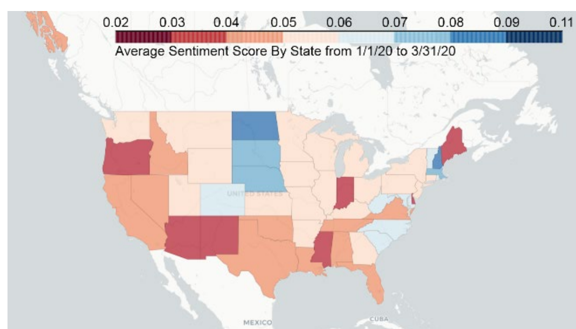

Figure 5. Heatmap of average tweet sentiment by state from 1/1/20 to 3/31/20

5. Discussion

We investigated the potential of Twitter data to predict COVID-19 cases and death cases in the US. We examined both same-day predictions and a lagged model that predicted counts two days into the future.

Our results indicated that the Twitter user's collective mood as represented by the sentiments of posts offered a high degree of accuracy (above 95% for all models) in predicting cases and death counts, both for the same-day and several days into the future. Sentiments of social media posts offer valuable information to public health officials for nowcasting and forecasting during a pandemic.

These results suggest that current social media sentiment can be used to identify the need for targeted information campaigns, creating or changing public health local and state policies, and overall targeting specific areas. Moreover, on days when official reporting can be slowed, such as weekends or holidays, social media sentiment can be used to predict cases and deaths and inform decision-making.

While this study relied on data from the United States, the results may extend to other countries. This can be valuable for countries with less developed reporting infrastructure or those that experience delays in reporting final official numbers. For example, it can take Germany several days before final COVID-19 cases are reported, and they are significantly underreported on the weekends [31]. Social media sentiment can be used to complement official numbers in these instances.

While several studies have modeled case and death counts of COVID-19 with the *volume* of disease-related social media posts [11], this is among the first studies to do so focusing on *sentiment* of posts. Despite the diverse sentiments of individual posters, the collective sentiments expressed on social media is sometimes considered a proxy of the public mood towards a particular topic. Shifts in mood expressed on social media occur in real-time in response to new information, new experiences, and new events. Just as these shifts have been shown to predict real-world outcomes like financial market movements [20], they likewise have value in modeling public health outcomes [28]. In doing so, we add to the growing literature investigating the application of sentiment analysis for combatting pandemics and outbreaks.

These results offer insights for future research. First, this analysis produced a single predictive model that can handle the variability of COVID-19 cases and deaths across all 50 states. Subsequent research should model sentiments at different levels of analysis, such as by state or by county. Furthermore, there is a need to investigate how the accuracy of these predictive models changes over time, which would require fitting many models over increasingly larger portions of the study period and would require model fits for each state.

This study primarily focused on the overall Twitter sentiments related to COVID-19. Future studies can investigate whether sentiment towards specific topics (e.g., mask-wearing, hand washing, CDC reports, etc.) has a greater influence on the accuracy of the model. To this end, combining topic modeling and sentiment analysis may provide more nuanced understandings of how social media sentiment might relate to case and death counts (e.g., does prevailing negative sentiment about mask-wearing predict increased case rate in a geographical area?).

While social media sentiment has been used to predict a range of real-world outcomes, the theoretical underpinnings of these findings are less developed. This is particularly important when we consider that social media platforms are not a representative cross-section of the United States. That is to say, it is incorrect to assert that the tweet corpus in this study represents public sentiment. Compared to the general public, Twitter users are younger, more educated, more demographic, and wealthier [32]. Despite this limited segment of the population, the sentiments of this group are nonetheless sufficient to accurately predict COVID-19 cases and deaths. Future research may better theorize connections between public expressions of sentiment on social media and real-world health outcomes. A promising theoretical framework is the Integrated Behavior Model (IBM) [33]. This theory of behavioral prediction asserts that attitudes, perceived norms, and personal agency together influence intention to engage in specific health behaviors. Discussions on Twitter may both embody the attitudes held by the public concerning behaviors or health issues, beliefs about others' behaviors and beliefs, and one's own control over a public health event. Applying IBM to tweets has yielded insightful public health insights in areas such as determinants of health behaviors towards human papillomavirus vaccination [34]. Such an approach applied to infectious disease pandemics may be useful in developing interventions to encourage protective behaviors by addressing social norms, increasing positive attitudes towards protective behaviors, and improving individual perceived control over their actions.

6. Conclusion and Limitations

In this study, we modeled COVID-19 case and death counts using the sentiment of Twitter posts. Using automated sentiment analysis, we show that the public sentiment of Twitter users offers a strong predictor of both same-day and future case and death rates. As public health officials and policymakers seek data that moves faster than the disease, there is perhaps no source faster than the real-time sentiments of people expressed on social media. This can be particularly useful in the early stages of a pandemic when there is little information available to public health officials and when disease testing lacks widespread availability. Furthermore, this data could be used for target interventions to address misinformation and promote specific behaviors.

This research expands the existing literature on the application of sentimental analysis on social media to assist in tracking and responding to outbreaks and pandemics [11, 23, 35]. Our findings also support previous research demonstrating that Twitter offers rich real-time information on opinions that can be used by public health authorities [22].

There are limitations to this research. First, the analysis includes Twitter posts from the first three months of the COVID-19 pandemic. While we show that sentiment is a useful predictor at the crucial beginning of a pandemic, the accuracy of the modeling may change as the pandemic persists. Second, the data was limited to English-language Twitter posts, which tend to represent a segment of the overall population. Expanding models to include a range of social media platforms and translating Tweets from different languages may increase the reliability and accuracy of these predictions. Third, the sentiment of tweets was geocoded to the state level. The application of these findings to smaller localized areas (e.g., region, county, or city) requires further study.

7. References

[1] *Data Modernization Initiative Basics: The Why, What, and How of Data Modernization,* Report Number 319521-T, Centers for Disease Control and Prevention, 2021. [Online]. Avaiable: https://www.cdc.gov/surveillance/pdfs/dmi_basics_external_audiences-March_2021.pdf.

[2] X. Ji, S. A. Chun, Z. Wei, and J. Geller, "Twitter sentiment classification for measuring public health concerns," *Social Network Analysis and Mining,* vol. 5, no. 1, 2015.

[3] G. Eysenbach, "Infodemiology: tracking flu-related searches on the web for syndromic surveillance," in Proceedings of the American Medical Informatics Association (AMIA) Annual Symposium, 2006, pp. 244-248.

[4] G. Eysenbach, "Infodemiology and Infoveillance," *American Journal of Preventive Medicine,* vol. 40, no. 5, 2011, pp. S154-S158.

[5] G. Eysenbach, "Infodemiology and Infoveillance: Framework for an Emerging Set of Public Health Informatics Methods to Analyze Search, Communication and Publication Behavior on the Internet," *Journal of Medical Internet Research,* vol. 11, no. 1, 2009, pp. e11.

[6] J. B. Colditz, K.-H. Chu, S. L. Emery, C. R. Larkin, A. E. James, J. Welling, and B. A. Primack, "Toward Real-Time Infoveillance of Twitter Health Messages," *American Journal of Public Health,* vol. 108, no. 8, 2018, pp. 1009-1014.

[7] A. Mavragani, "Tracking COVID-19 in Europe: Infodemiology Approach," *JMIR Public Health and Surveillance,* vol. 6, no. 2, 2020, pp. e18941.

[8] J. Turiel, D. Fernandez-Reyes, and T. Aste, "Wisdom of Crowds Detects COVID-19 Severity Ahead of Officially Available Data," *arXiv: 2004.04125v4 [physics.soc-ph]*, Jun. 2020.

[9] J. Li, Q. Xu, R. Cuomo, V. Purushothaman, and T. Mackey, "Data Mining and Content Analysis of the Chinese Social Media Platform Weibo During the Early COVID-19 Outbreak: Retrospective Observational Infoveillance Study," *JMIR Public Health and Surveillance,* vol. 6, no. 2, 2020, pp. e18700.

[10] C. Hutto, and E. Gilbert, "Vader: A parsimonious rule-based model for sentiment analysis of social media text," in Proceedings of the Eighth International AAAI Conference on Web and Social Media (ICWSM), Ann Arbor, MI, 2014.

[11] A. H. Alamoodi, B. B. Zaidan, A. A. Zaidan, O. S. Albahri, K. I. Mohammed, R. Q. Malik, E. M. Almahdi, M. A. Chyad, Z. Tareq, A. S. Albahri, H. Hameed, and M. Alaa, "Sentiment analysis and its applications in fighting COVID-19 and infectious diseases: A systematic review," *Expert Systems with Applications,* vol. 167, 2021, pp. 114155.

[12] H. Achrekar, A. Gandhe, R. Lazarus, S.-H. Yu, and B. Liu, "Predicting flu trends using twitter data," in IEEE Conference on Computer Communications Workshops, 2011, pp. 702-707.

[13] S. Masri, J. Jia, C. Li, G. Zhou, M.-C. Lee, G. Yan, and J. Wu, "Use of Twitter data to improve Zika virus surveillance in the United States during the 2016 epidemic," *BMC Public Health,* vol. 19, no. 1, 2019, pp. 761.

[14] E. X. Zhang, Y. Yang, R. Di Shang, J. J. P. Simons, B. K. Quek, X. F. Yin, W. See, O. S. H. Oh, K. S. T. Nandar, V. R. Y. Ling, P. P. Chan, Z. Wang, R. S. M. Goh, L. James, and J. S. H. Tey, "Leveraging social networking sites for disease surveillance and public sensing: the case of the 2013 avian influenza A(H7N9) outbreak in China," *Western Pacific Surveillance and Response Journal,* vol. 6, no. 2, 2015, pp. 66-72.

[15] H. Rosenberg, S. Syed, and S. Rezaie, "The Twitter pandemic: The critical role of Twitter in the dissemination of medical information and misinformation during the COVID-19 pandemic," *Canadian Journal of Emergency Medicine,* vol. 22, no. 4, 2020, pp. 418-421.

[16] A. Hussain, A. Tahir, Z. Hussain, Z. Sheikh, M. Gogate, K. Dashtipour, A. Ali, and A. Sheikh, "Artificial Intelligence–Enabled Analysis of Public Attitudes on Facebook and Twitter Toward COVID-19 Vaccines in the United Kingdom and the United States: Observational Study," *Journal of Medical Internet Research,* vol. 23, no. 4, 2021, pp. e26627.

[17] E. Gharavi, N. Nazemi, and F. Dadgostari, "Early Outbreak Detection for Proactive Crisis Management Using Twitter Data: COVID-19 a

Case Study in the US," arXiv: 2005.00475v1 [cs.SI], May 2020.
[18] R. Kouzy, J. Abi Jaoude, A. Kraitem, M. B. El Alam, B. Karam, E. Adib, J. Zarka, C. Traboulsi, E. W. Akl, and K. Baddour, "Coronavirus Goes Viral: Quantifying the COVID-19 Misinformation Epidemic on Twitter," *Cureus,* vol. 12, no. 3, 2020, pp. e7255-e7255.
[19] P. Chauhan, N. Sharma, and G. Sikka, "The emergence of social media data and sentiment analysis in election prediction," *Journal of Ambient Intelligence and Humanized Computing,* vol. 12, no. 2, 2021, pp. 2601-2627.
[20] J. Bollen, H. Mao, and X. Zeng, "Twitter mood predicts the stock market," *Journal of Computational Science,* vol. 2, no. 1, 2011, pp. 1-8.
[21] S. Liu, B. Chen, and A. Kuo, "Monitoring Physical Activity Levels Using Twitter Data: Infodemiology Study," *Journal of Medical Internet Research,* vol. 21, no. 6, 2019, pp. e12394.
[22] C. Chew, and G. Eysenbach, "Pandemics in the Age of Twitter: Content Analysis of Tweets during the 2009 H1N1 Outbreak," *PLOS ONE,* vol. 5, no. 11, 2010, pp. e14118.
[23] R. Singh, and R. Singh, "Applications of sentiment analysis and machine learning techniques in disease outbreak prediction–A review," *Materials Today: Proceedings*, 2021.
[24] A. D. Dubey, "Twitter sentiment analysis during COVID19 outbreak," (April 9, 2020). Available at SSRN: http://dx.doi.org/10.2139/ssrn.3572023.
[25] S. Boon-Itt, and Y. Skunkan, "Public Perception of the COVID-19 Pandemic on Twitter: Sentiment Analysis and Topic Modeling Study," *JMIR Public Health Surveillance,* vol. 6, no. 4, 2020, pp. e21978.
[26] A. Gupta, and R. Katarya, "Social media based surveillance systems for healthcare using machine learning: A systematic review," *Journal of Biomedical Informatics,* vol. 108, 2020, pp. 103500.
[27] "Coronavirus (COVID-19) Data in the United States," The New York Times, 2021. [Online]. Avaiable: https://github.com/nytimes/covid-19-data
[28] D. Valdez, M. Ten Thij, K. Bathina, L. A. Rutter, and J. Bollen, "Social Media Insights Into US Mental Health During the COVID-19 Pandemic: Longitudinal Analysis of Twitter Data," *Journal of Medical Internet research,* vol. 22, no. 12, 2020, pp. e21418.
[29] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society: Series B (Methodological),* vol. 58, no. 1, 1996, pp. 267-288.
[30] A. E. Hoerl, and R. W. Kennard, "Ridge regression: Biased estimation for nonorthogonal problems," *Technometrics,* vol. 12, no. 1, 1970, pp. 55-67.
[31] B. Skiera, L. Jürgensmeier, K. Stowe, and I. Gurevych, "How to best predict the daily number of new infections of COVID-19," arXiv: 2004.03937v1 [cs.SI], Apr. 2020.
[32] S. Wojcik, and A. Hughes, "Sizing Up Twitter Users," *Pew Research Center*, 2018.
[33] M. S. Hagger, and N. L. Chatzisarantis, "Integrating the theory of planned behaviour and self-determination theory in health behaviour: A meta-analysis," *British Journal of Health Psychology,* vol. 14, no. 2, 2009, pp. 275-302.
[34] H. Zhang, C. Wheldon, A. G. Dunn, C. Tao, J. Huo, R. Zhang, M. Prosperi, Y. Guo, and J. Bian, "Mining Twitter to assess the determinants of health behavior toward human papillomavirus vaccination in the United States," *Journal of the American Medical Informatics Association,* vol. 27, no. 2, 2020, pp. 225-235.
[35] R. Singh, and R. Singh, "Machine Learning Based Twitter Data Mining to Analyse Sentiments of Tweets Allied to COVID-19 Epidemic & Its Patterns," *International Journal of Information Technology and Electrical Engineering,* vol. 9, no. 2, 2020, pp. 11-22.